Queen's University Belfast

Introduction to Support Vector Machine

Yang Hua

Mar. 2017

Outline

- Ice breaker
- Geometric viewpoint of (linear) SVM
- Statistical learning viewpoint of (linear) SVM
- Non-linear SVM: Kernel tricks
- Engineering viewpoint of SVM: practical guide
- Summary

Ice breaker

- Data

	Height (m)	Weight (kg)	Gender
1	1.76	73.7	Male
2	1.71	75.1	Male
3	1.82	80.0	Male
4	1.64	60.1	Female
5	1.55	45.6	Female
6	1.67	52.5	Female

- Q1: How to estimate body type?
  - Data collection
  - Mathematical modeling (Variables, Constraints, Objective function …)

$$\mathrm{BMI} = \frac{\mathrm{mass\ (kg)}}{(\mathrm{height(m)})^2}$$

  - Implementation & evaluation

Ice breaker

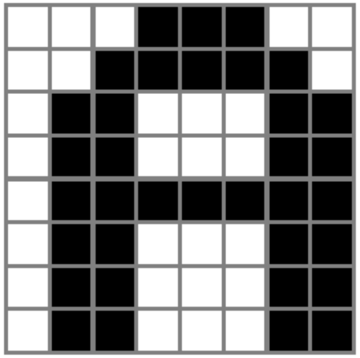
- Data

	Height (m)	Weight (kg)	Gender
1	1.76	73.7	Male
2	1.71	75.1	Male
3	1.82	80.0	Male
4	1.64	60.1	Female
5	1.55	45.6	Female
6	1.67	52.5	Female

- Q2: How to guess gender?
  - How could we use these data to predict new person’s gender?

Ice breaker

- More harder problem: handwriting recognition



  - It is difficult to model it mathematically

```
if (I[0,5]<128) & (I[0,6] > 192) & (I[0,7] < 128):
  return 'A'
elif (I[7,7]<50) & (I[6,3]) != 0:
  return 'Q'
else:
  print "I don't know this letter."
```

Ice breaker

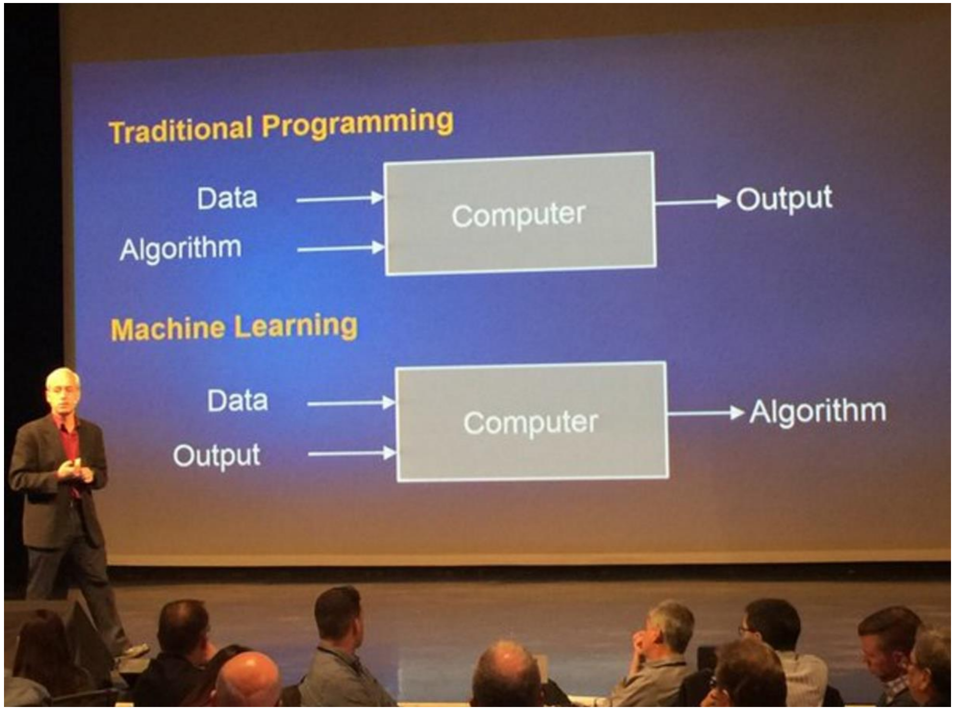
- Human vs. Machine
  - Human is intelligent & wise, but human is “lazy” & expensive
  - Machine is powerful and cheap, but machine is “stiffness”
- How do we teach machine to learn? To let machine with intelligence?
  - **Machine learning is one approach to Artificial Intelligence (AI)**

Ice breaker

- Machine learning – experts' perspectives
  - **Tom M. Mitchell**: A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.
  - **Andrew Ng**.: Machine Learning is the science of getting computers to learn, without being explicitly programmed.
  - **Christoph Lampert**: The science of automatic systems that draw conclusions from empirical data.

Ice breaker

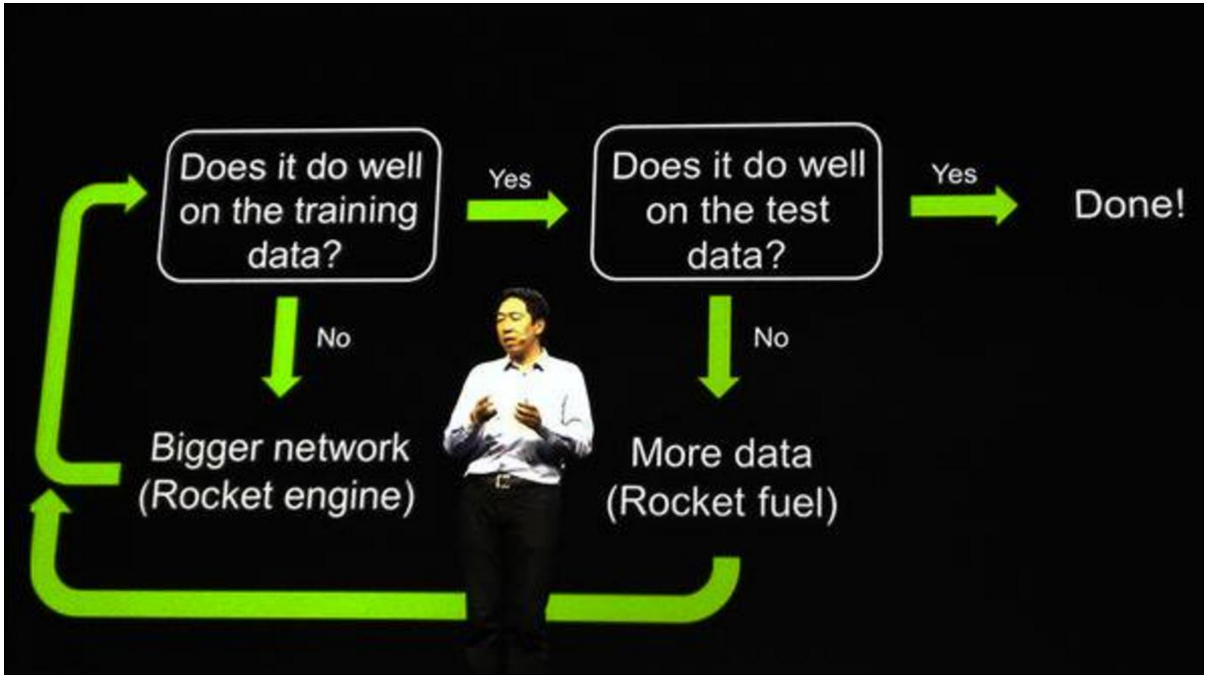
- Machine learning – "Informal" viewpoint
  - Machine learning = Statistics (modeling) + Optimization (solver)
  - We solve the problem by means of data without explicitly knowing the true model
  - We don't model/program a solution to the specific problem



* Image courtesy of Dr. Pedro Domingos, who is a professor at UW

Ice breaker

- Why we need to learn some machine learning theory?
  - Applications cannot be carried out by simply using a black box.
  - What is needed:
    - choice of representation (inputs, outputs)
    - choice of learning model
    - analysis of evaluation results



Outline

- Ice breaker
- Geometric viewpoint of (linear) SVM
- Statistical learning viewpoint of (linear) SVM
- Non-linear SVM: Kernel tricks
- Engineering viewpoint of SVM: practical guide
- Summary

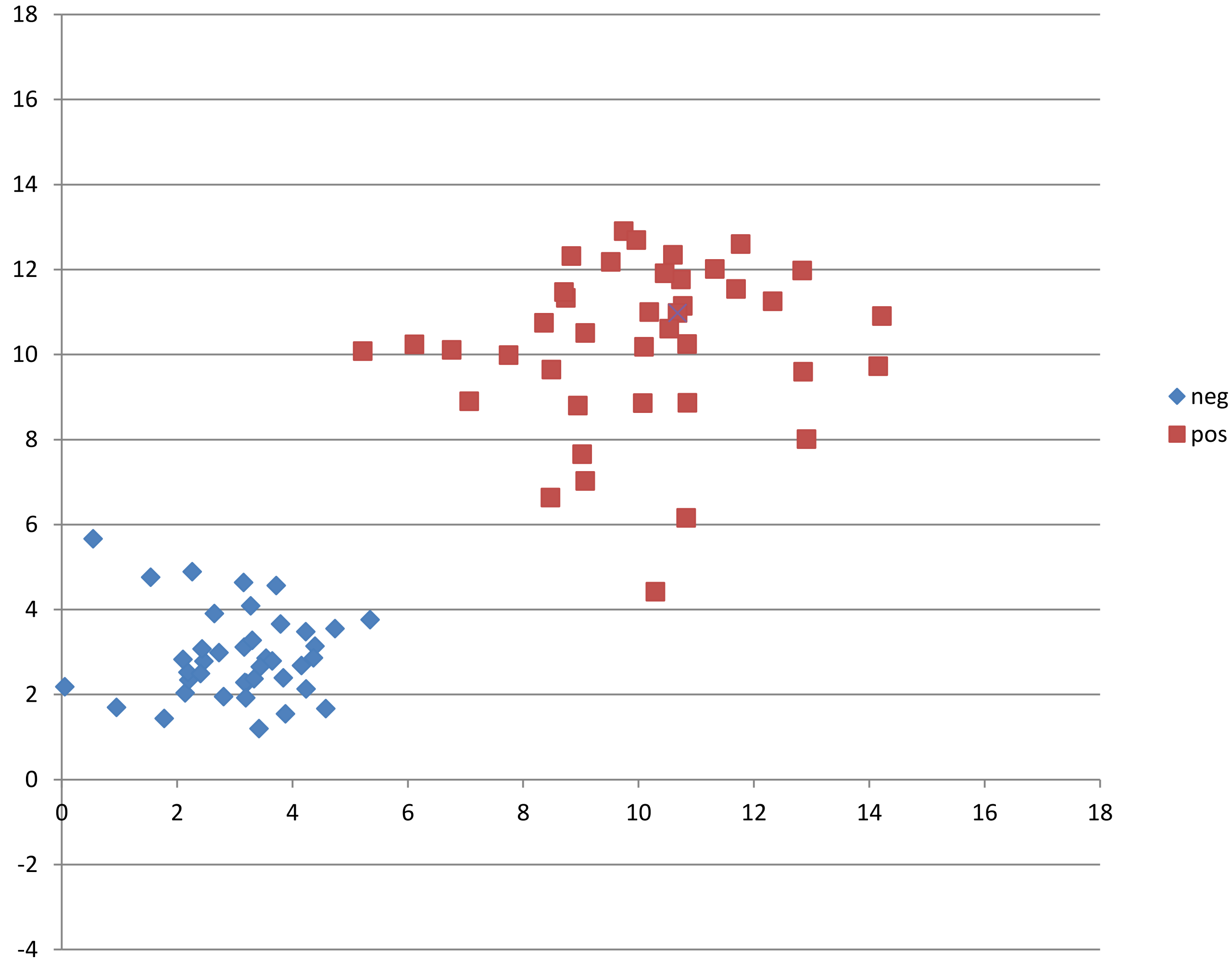
Geometric viewpoint of (linear) SVM

- Premise: terms
  - Supervised learning vs. unsupervised learning
  - Classification vs. Regression
  - Training data vs. testing data
- Premise: notations
  - Data points: $X = \{x_1, x_2, \ldots x_m\}, x_i \in \mathcal{R}^n$, e.g. $x_i \in \mathcal{R}^2$ (height, weight)
  - Class labels: $Y = \{y_1, y_2, \ldots y_m\}, y_i \in$ {+1, -1}, e.g. male(+1), female(-1)
  - Goal: From training data (X with Y), we can find optimal w and b, which can be used for predicting new testing data (x), i.e.,

$$g(x) = sign\, f(x), \text{ where } f(x) = w^T x + b$$

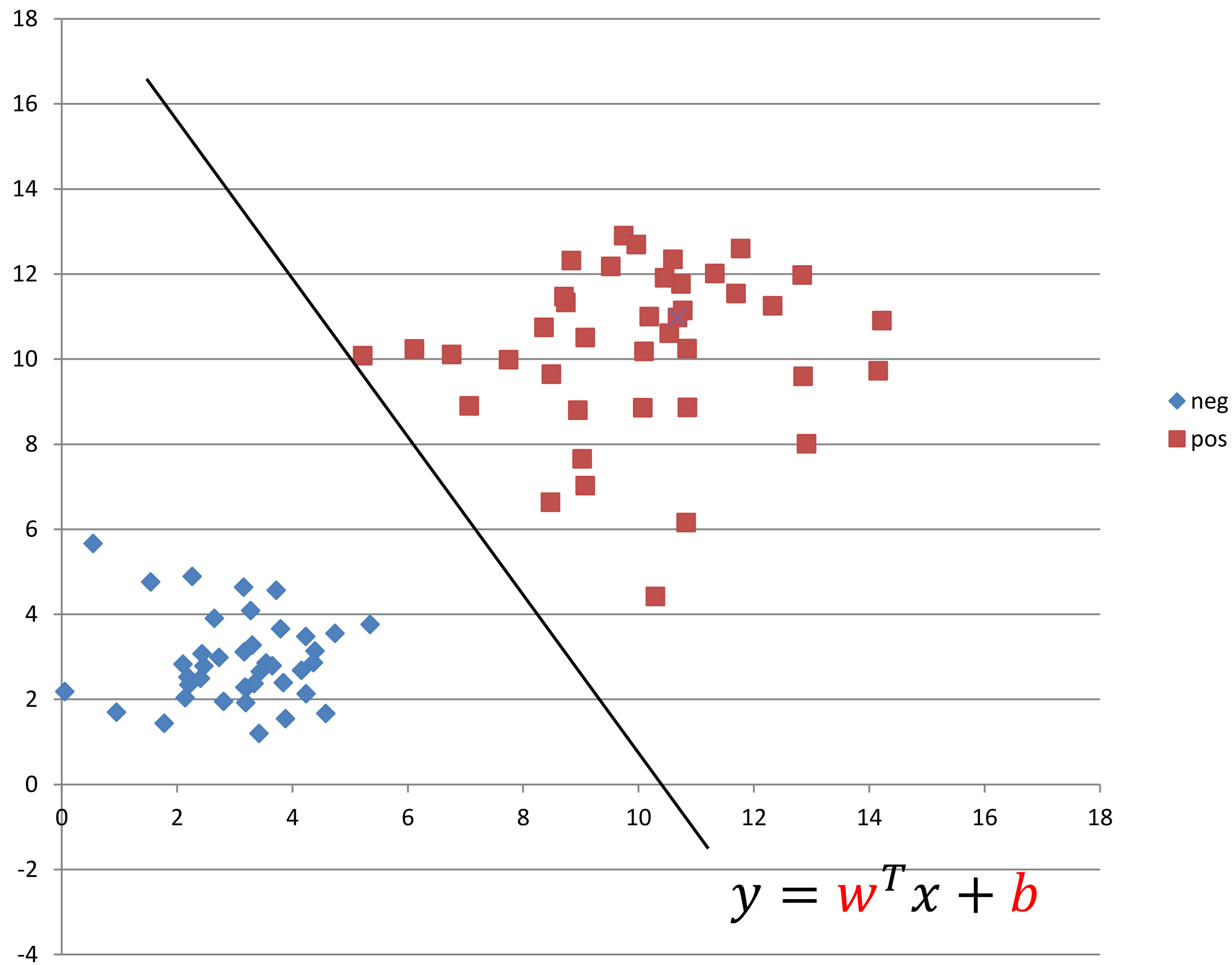
Geometric viewpoint of (linear) SVM

- Toy data



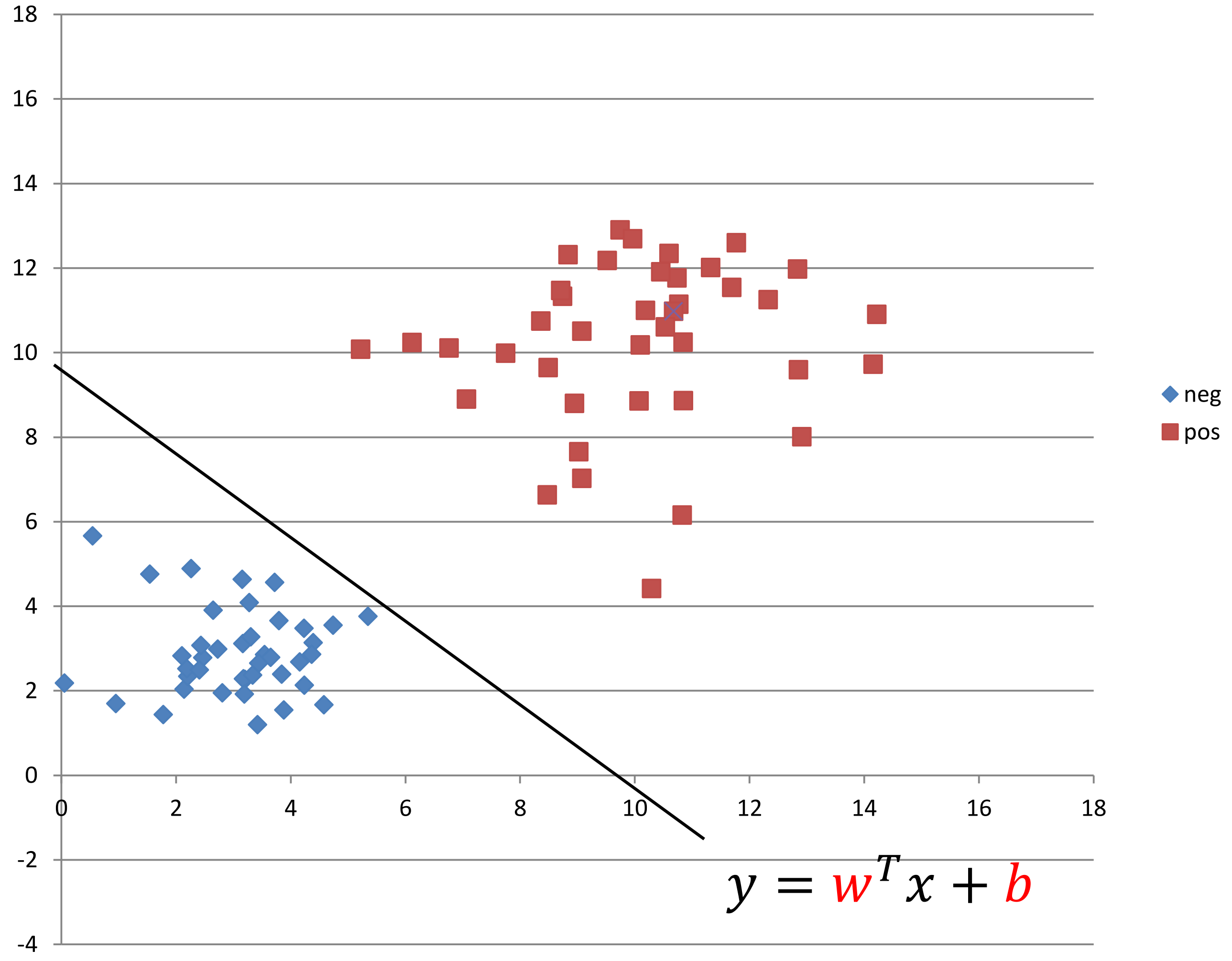
Geometric viewpoint of (linear) SVM

- Toy data



Geometric viewpoint of (linear) SVM

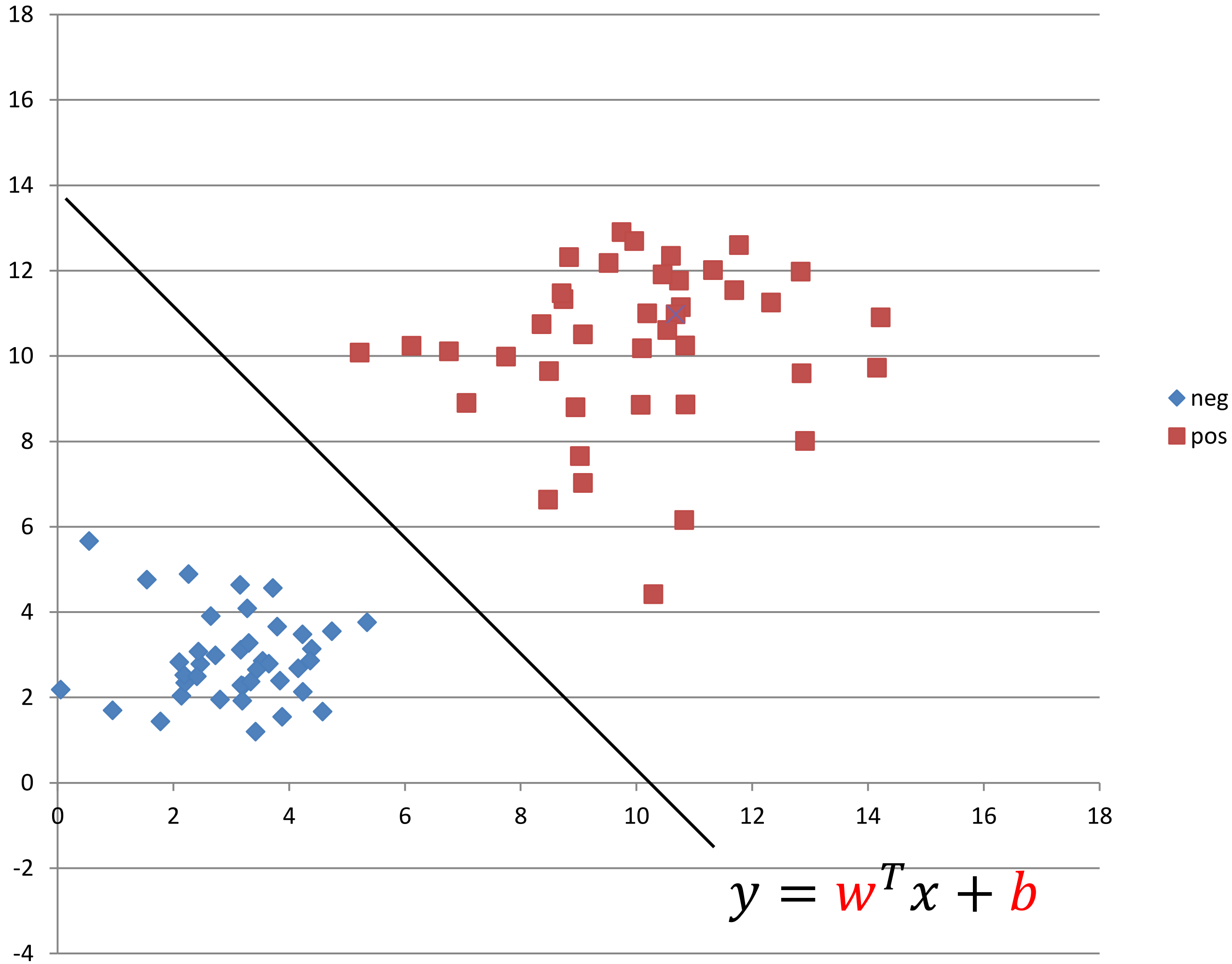
- Toy data



$$y = w^T x + b$$

Geometric viewpoint of (linear) SVM

- Toy data



Geometric viewpoint of (linear) SVM

- Goal: From training data (X with Y), we can find optimal w and b, which can be used for predicting new testing data (x), i.e.,

$$g(x) = sign\, f(x), \text{ where } f(x) = w^T x + b$$

- Method 1: linear algebra

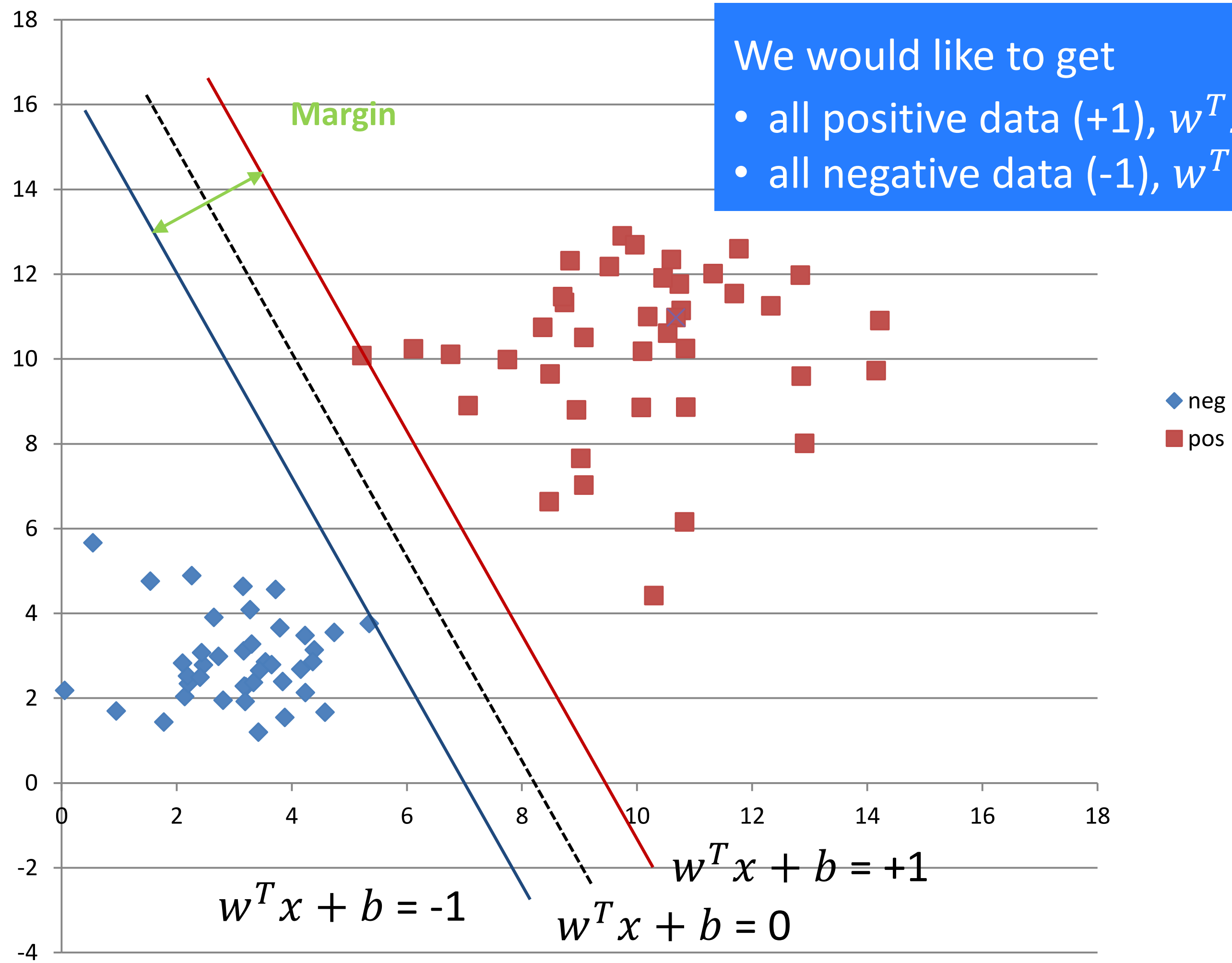
We want $f(x_i) > 0$ for $y_i = +1$ and $f(x_i) < 0$ for $y_i = -1$.
Let's just try $f(x_i) = y_i$ and solve

$$\begin{aligned} & w^t \boldsymbol{X} && = \boldsymbol{y} \\ \Rightarrow\quad & w^t \boldsymbol{X}\boldsymbol{X}^t && = \boldsymbol{y}\boldsymbol{X}^t \\ \Rightarrow\quad & w^t && = \underbrace{\boldsymbol{y}\boldsymbol{X}^t}_{1\times d} \underbrace{(\boldsymbol{X}\boldsymbol{X}^t)^{-1}}_{d\times d} \end{aligned}$$

  - Actually, it is the least square solution to let all positive data points to fit to $w^T x + b$ = +1 and all negative data points to fit to $w^T x + b$ = -1.

Geometric viewpoint of (linear) SVM

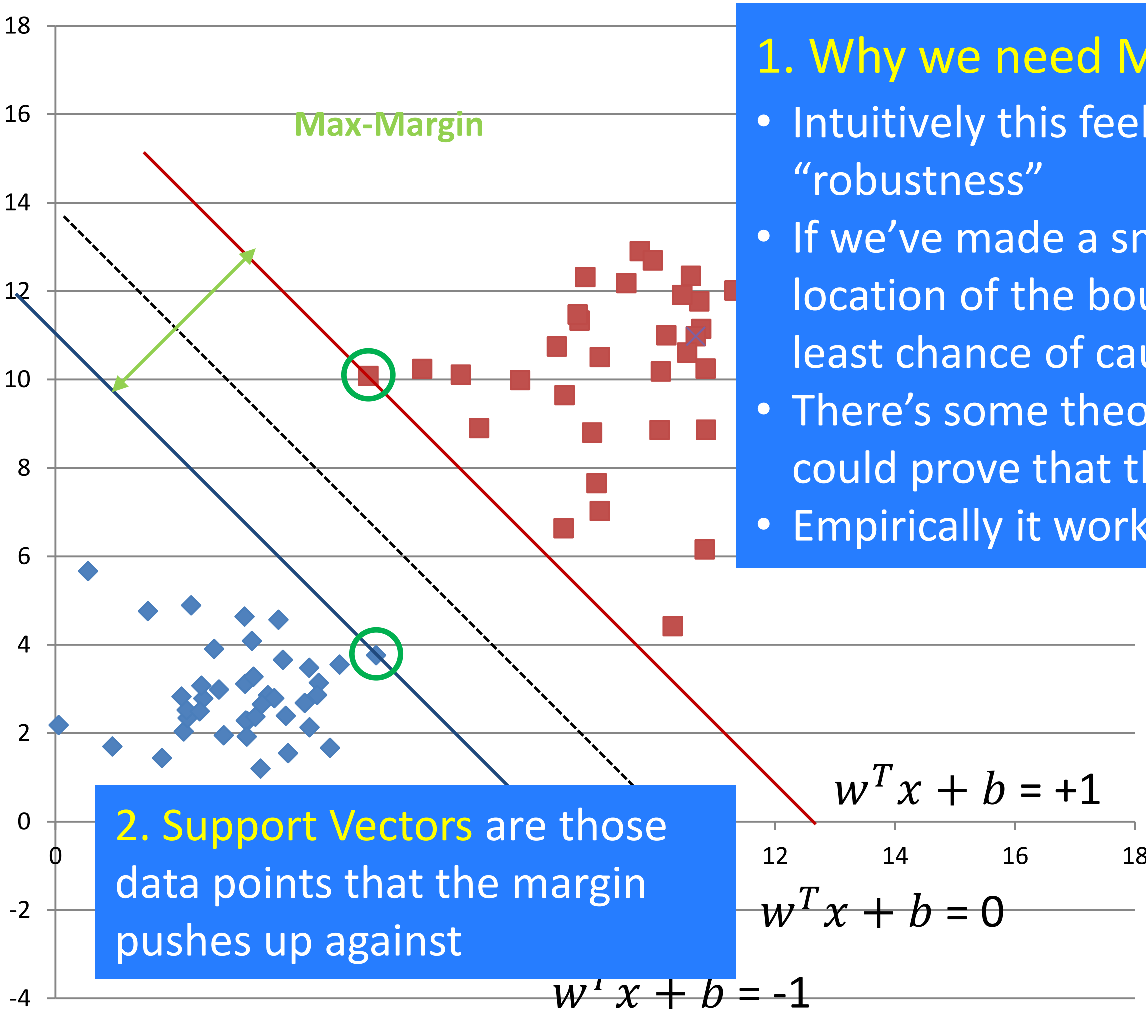
- Method 2: Margin



We would like to get

- all positive data (+1), $w^T x + b$ >= +1
- all negative data (-1), $w^T x + b$ <= -1

- Method 3: Maximum Margin



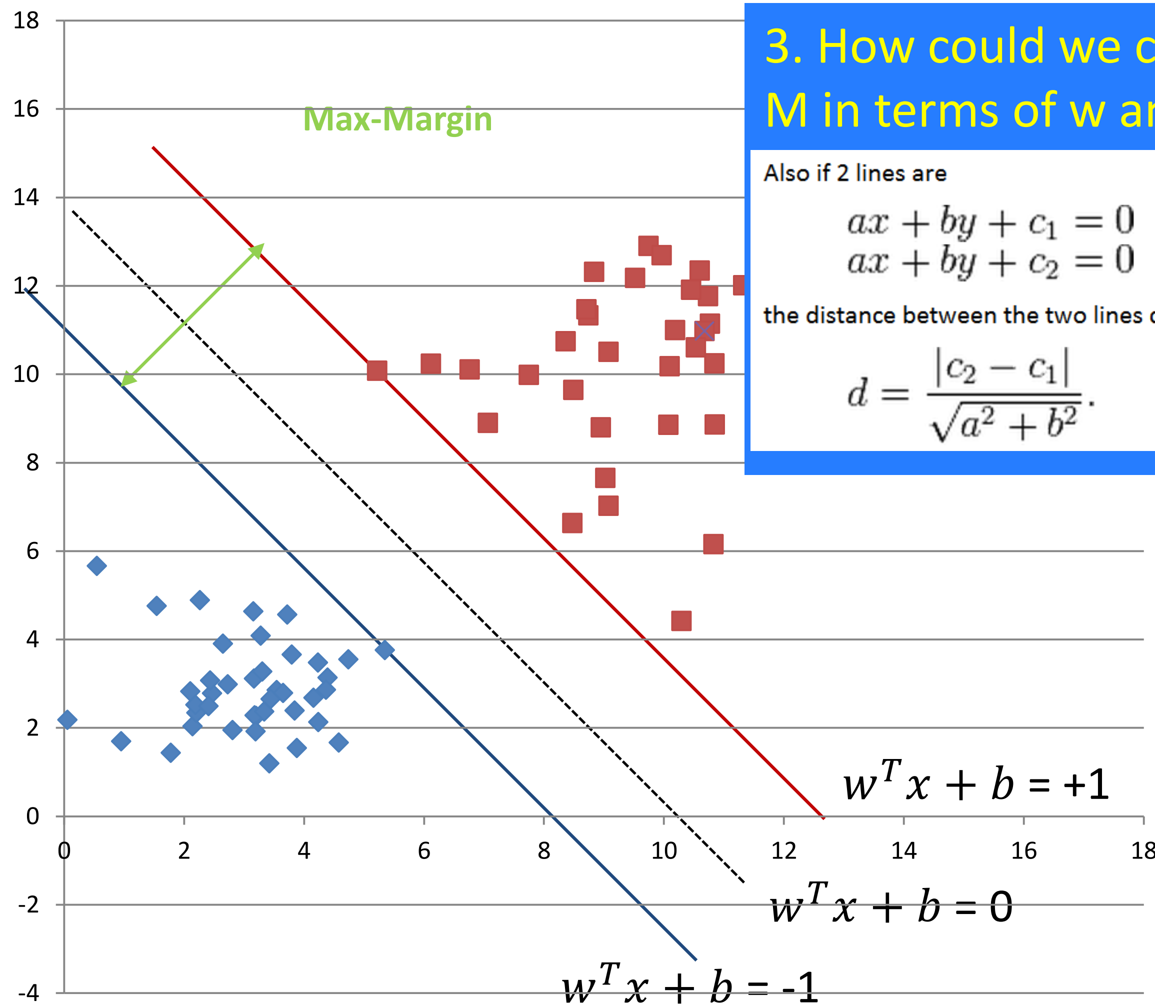
1. Why we need Maximum-Margin?

- Intuitively this feels safest and maximizes “robustness”
- If we’ve made a small error in the location of the boundary, this gives us least chance of causing a misclassification
- There’s some theory (using VC dimension) could prove that this is a good thing
- Empirically it works very well

2. Support Vectors are those data points that the margin pushes up against

Geometric viewpoint of (linear) SVM

- Method 3: Maximum Margin



3. How could we compute Max-Margin M in terms of w and b?

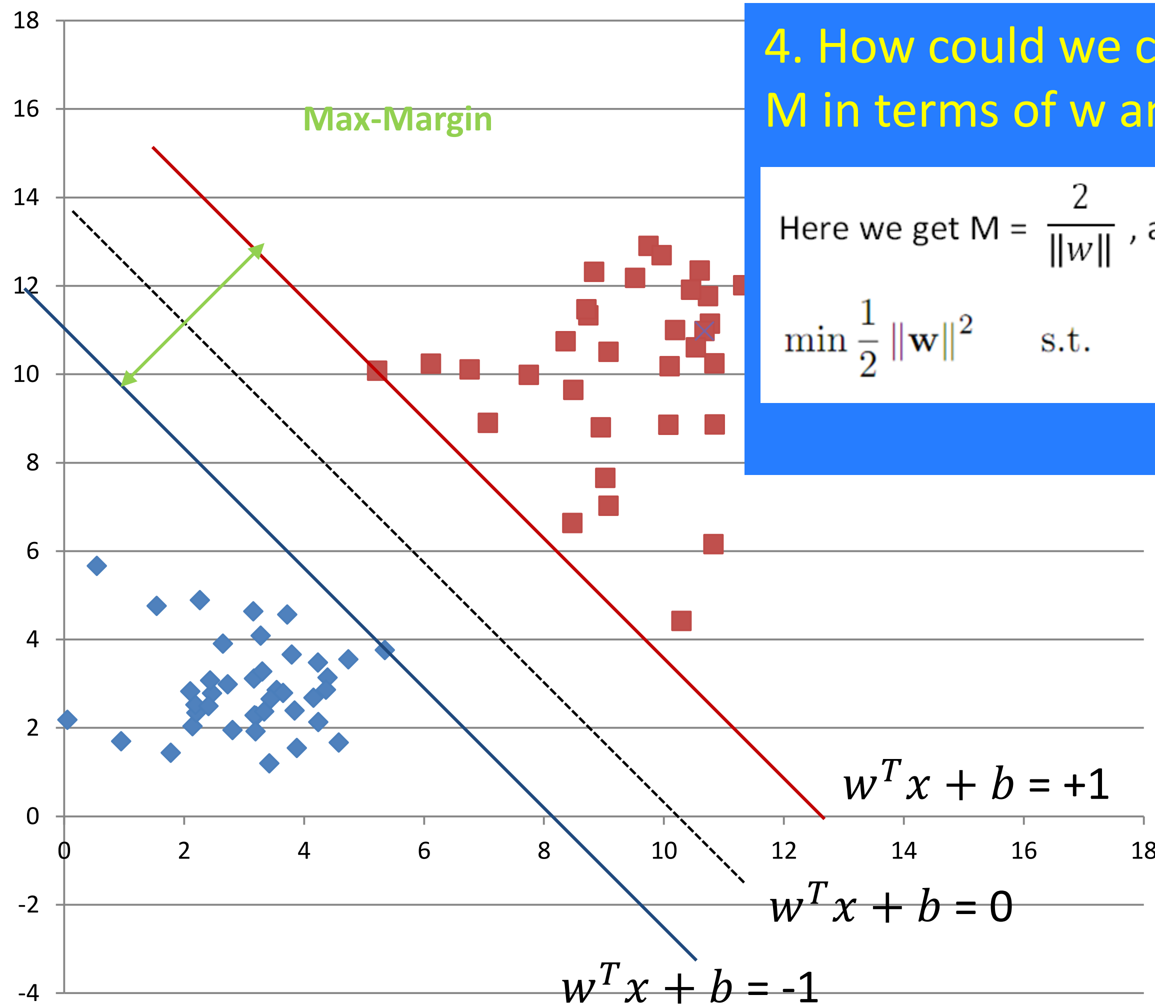
Also if 2 lines are

$$ax + by + c_1 = 0$$
$$ax + by + c_2 = 0$$

the distance between the two lines can be formulated by the following formula:

$$d = \frac{|c_2 - c_1|}{\sqrt{a^2 + b^2}}.$$

- Method 3: Maximum Margin



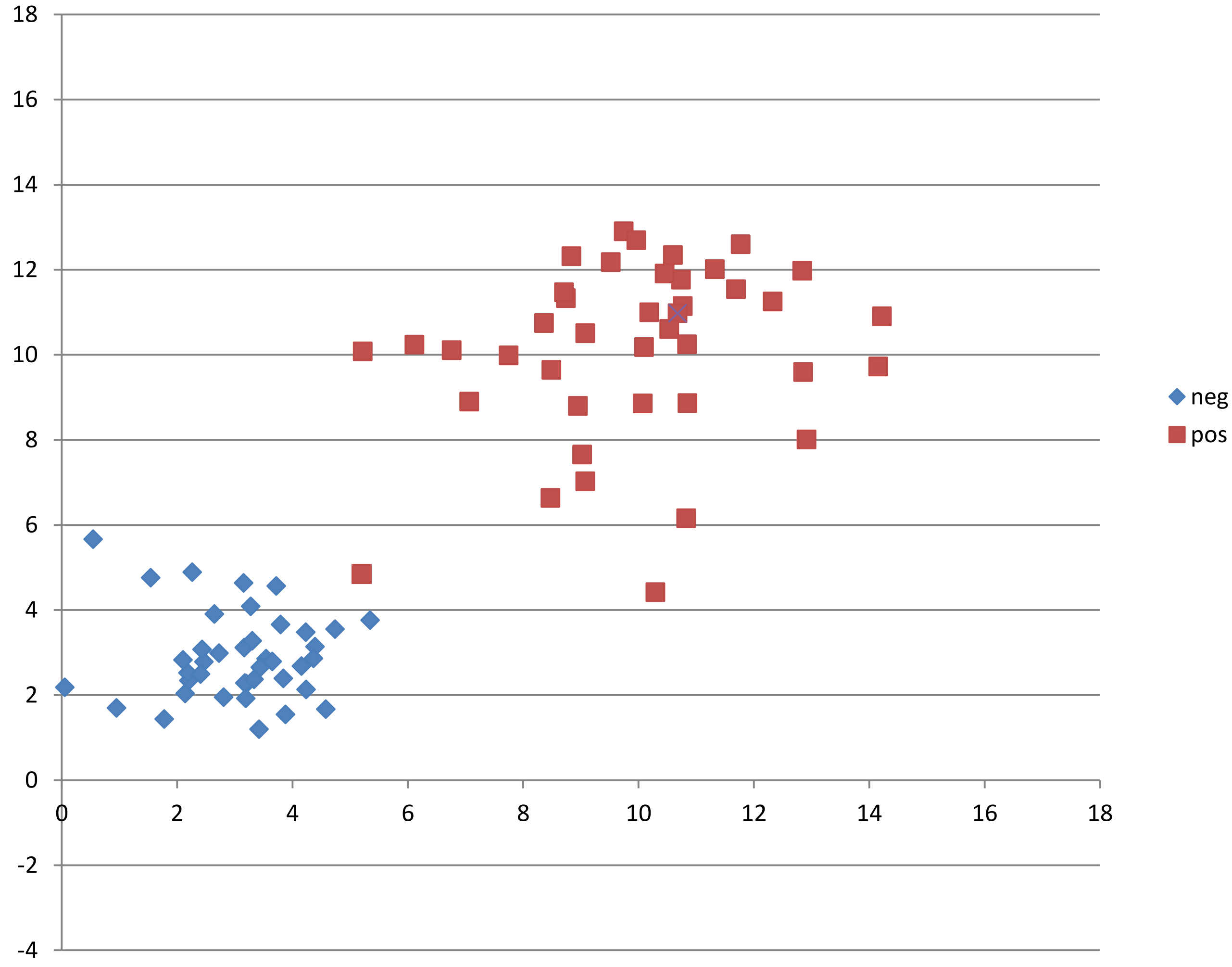
4. How could we compute Max-Margin M in terms of w and b?

Here we get M = $\frac{2}{\|w\|}$, and maximizing M equals to

$$\min \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{s.t.} \quad y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0 \;\; \forall_i$$

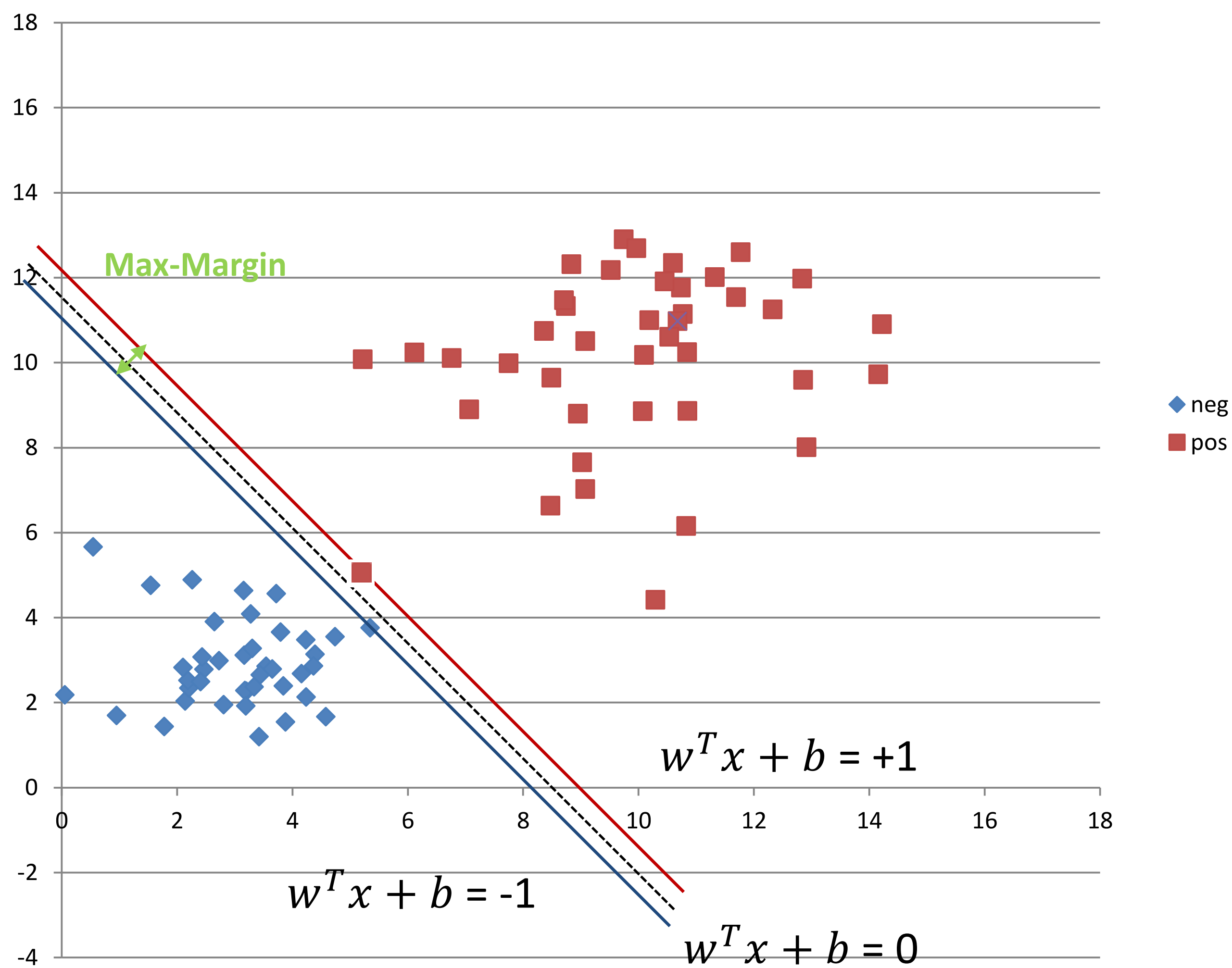
Geometric viewpoint of (linear) SVM

- How to solve such problem?



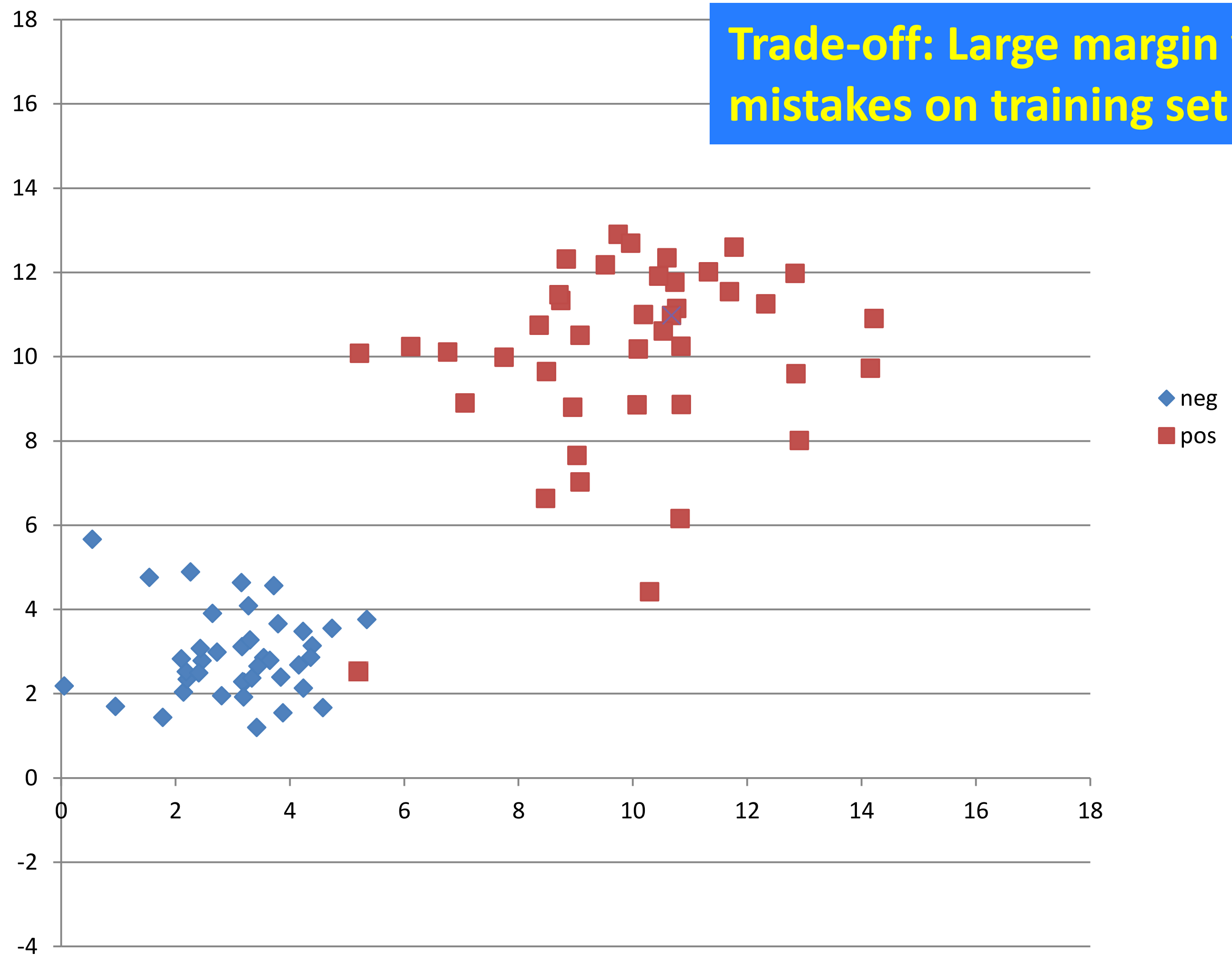
Geometric viewpoint of (linear) SVM

- Is Max-Margin still ok?



Geometric viewpoint of (linear) SVM

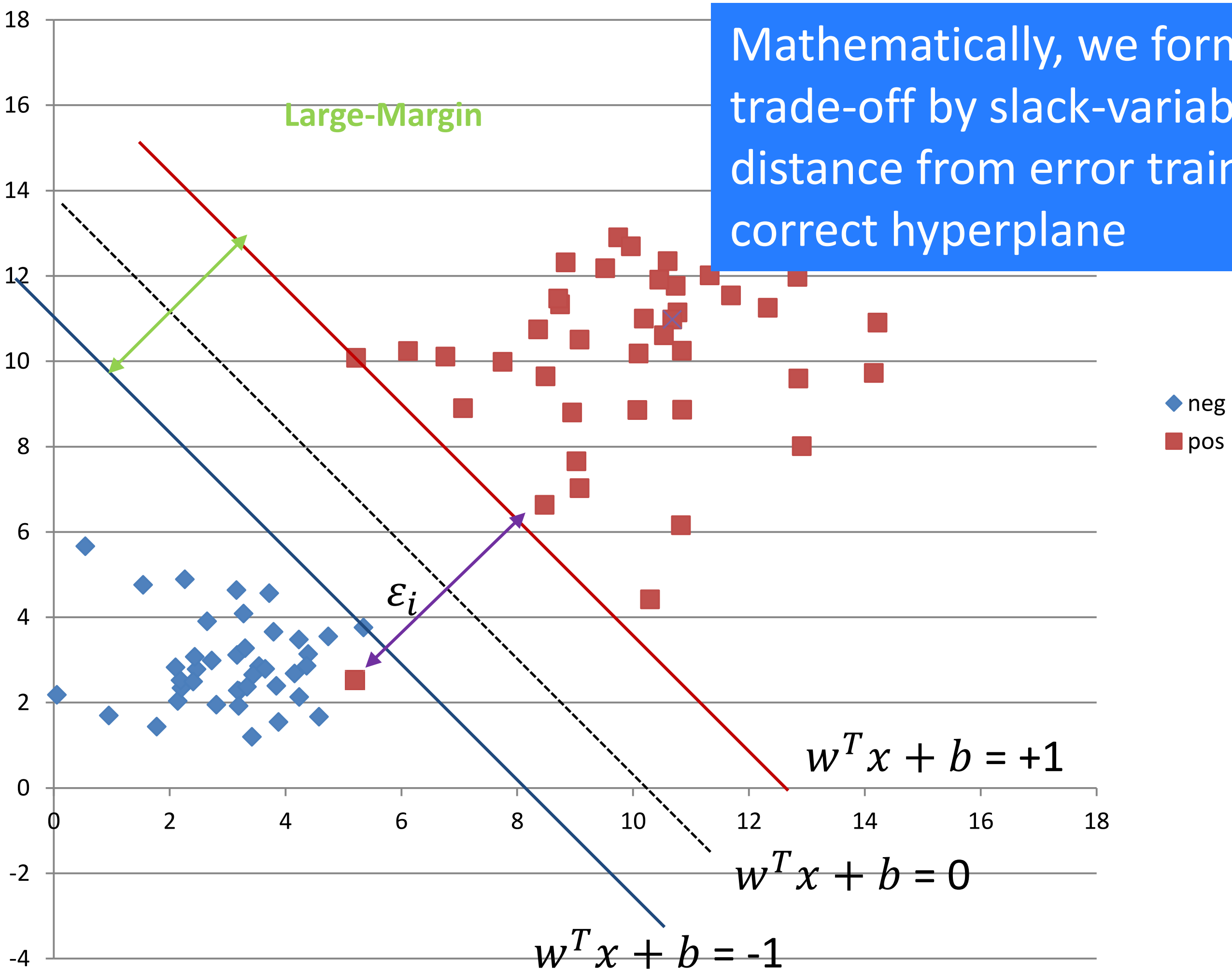
- And this?



Trade-off: Large margin vs. few mistakes on training set

Geometric viewpoint of (linear) SVM

- Method 4: Soft Margin



Mathematically, we formulate the trade-off by slack-variables ε_i - the distance from error training data to correct hyperplane

- Method 4: Soft Margin
  - So we get new form for trading-off large margin and few mistake training data

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{L}\xi_i$$

$$\text{s.t.} \qquad y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i \geq 0 \;\; \forall_i$$

$$\xi_i \geq 0 \; \forall_i$$

  - Discussion
    - We can fulfill *every* constraint by choosing ξ_i large enough.
    - The larger ξ_i, the larger the objective (that we try to minimize)
    - C is a *regularization*/trade-off parameter:
      - small $C \rightarrow$ constraints are easily ignored
      - large $C \rightarrow$ constraints are hard to ignore
      - $C = \infty \rightarrow$ hard margin case $\rightarrow$ no errors on training set
    - Note: The problem is still convex and efficiently solvable.

Outline

- Ice breaker
- Geometric viewpoint of (linear) SVM
- Statistical learning viewpoint of (linear) SVM
- Non-linear SVM: Kernel tricks
- Engineering viewpoint of SVM: practical guide
- Summary

Statistical learning viewpoint of (linear) SVM

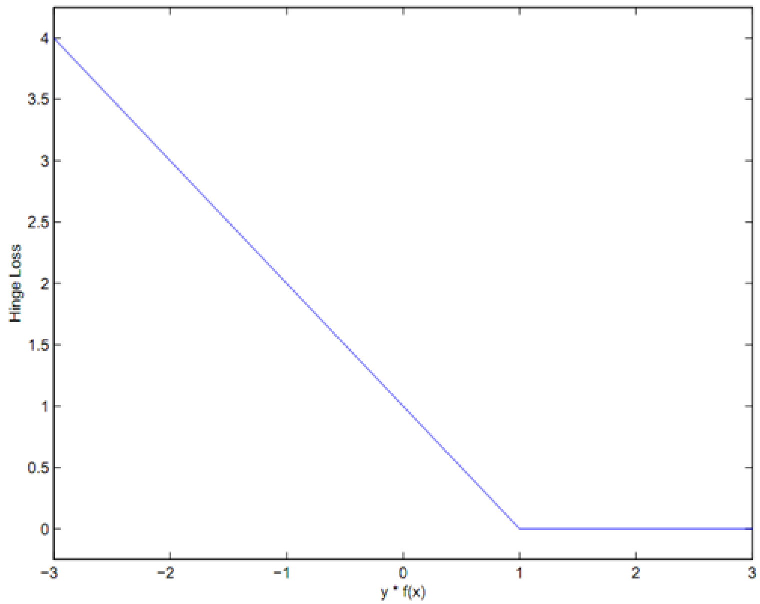
- Loss function: quantifies our unhappiness with the scores across the training data.

The classical SVM arises by considering the specific loss function

$$V(f(x,y)) \equiv (1 - yf(x))_+,$$

where

$$(k)_+ \equiv \max(k, 0).$$



The hinge loss

Statistical learning viewpoint of (linear) SVM

- Regularization
  - w is not unique, when it makes the loss become zero or small value.
  - Example

$$x = [1, 1, 1, 1]$$

$$w_1 = [1, 0, 0, 0]$$
$$w_2 = [0.25, 0.25, 0.25, 0.25]$$

$$w_1^T x = w_2^T x = 1$$

  - Why we prefer $w2$ to $w1$?
    - $w2$ is more general, i.e., use as much feature as possible. Otherwise, it will be easily overfitting the training dataset.
  - SVM uses L2 regularization techniques

Statistical learning viewpoint of (linear) SVM

- Connection between two viewpoints

Geometrical viewpoint

$$\min \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{L}\xi_i$$

$$\text{s.t.} \quad y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i \geq 0 \quad \forall_i$$

$$\xi_i \geq 0 \ \forall_i$$

$$\stackrel{?}{=}$$

Statistical learning viewpoint

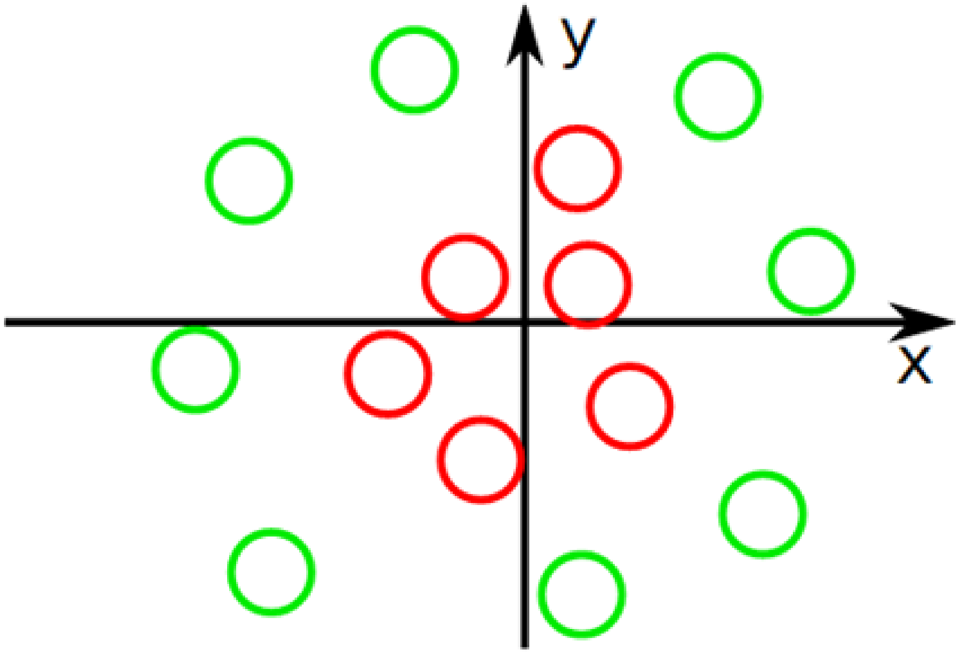
$$\underset{f \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(1 - y_i f(x_i))_+ + \lambda||f||^2_{\mathcal{H}}.$$

Outline

- Ice breaker
- Geometric viewpoint of (linear) SVM
- Statistical learning viewpoint of (linear) SVM
- Non-linear SVM: Kernel tricks
- Engineering viewpoint of SVM: practical guide
- Summary

Non-linear SVM: Kernel tricks

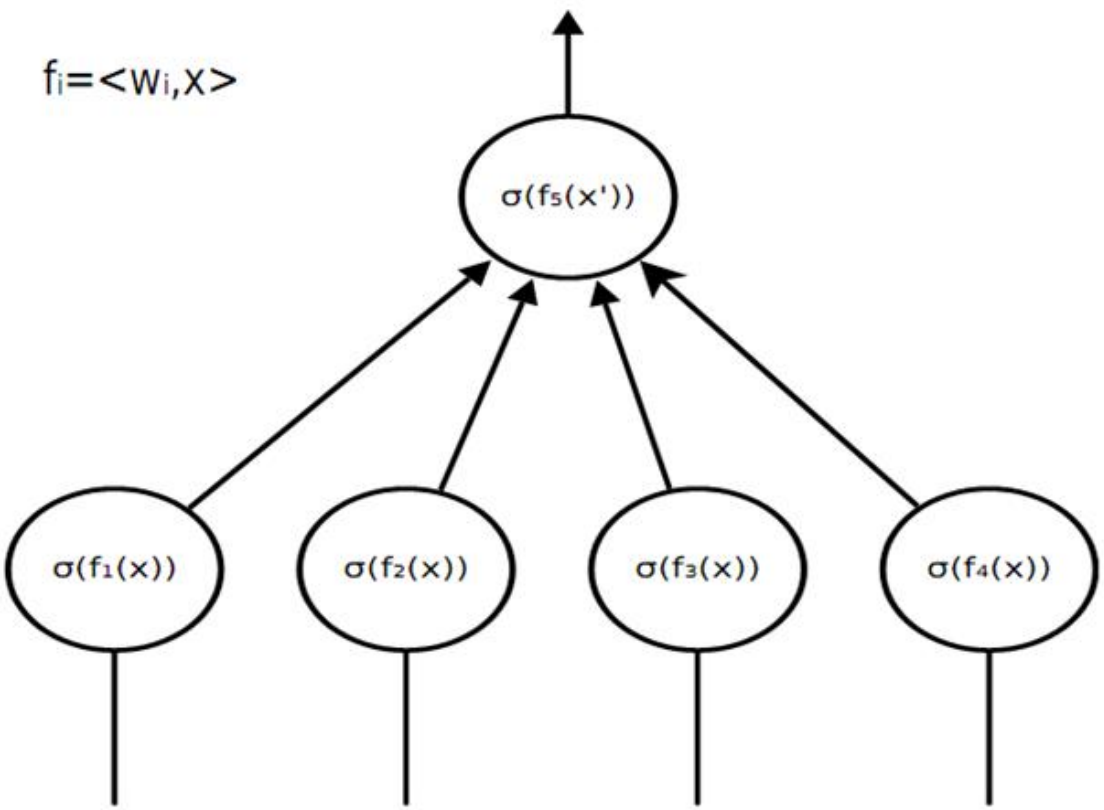
- What is the best soft-margin w for this dataset?



- None. We need something non-linear!

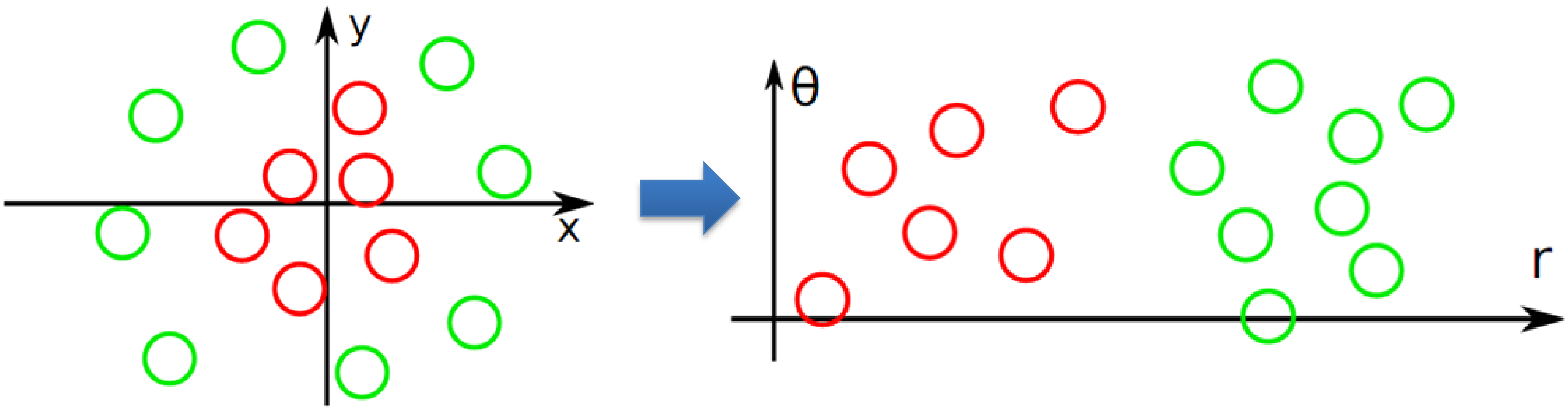
Non-linear SVM: Kernel tricks

- Method 1: Use classifier outputs as input to other classifier
  - Multiplayer Perceptron (a.k.a., (Artificial) Neural Network)
  - Boosting, Decision Trees, Random Forests



Non-linear SVM: Kernel tricks

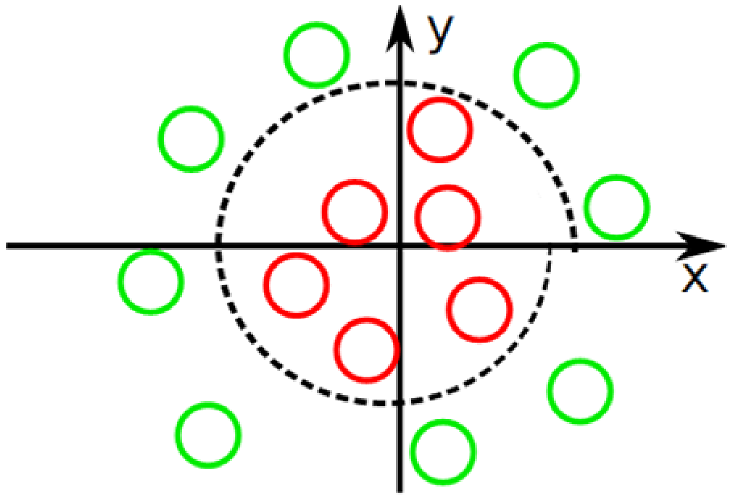
- Method 2: Preprocess the data

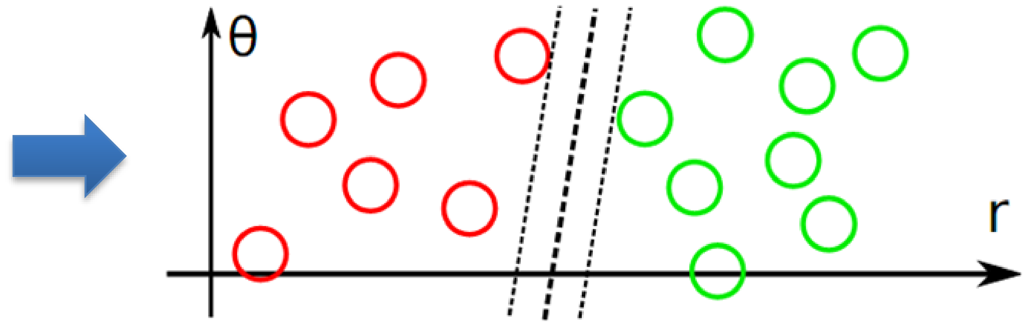


Cartesian coordinates

polar coordinates

- Method 2: Preprocess the data





Non-linear separation
in Cartesian space

Linear separation
in polar space

Non-linear SVM: Kernel tricks

- Does this operation always work?
  - Yes, if we do it right.

Lemma

Let $(x_i)_{i=1,\ldots,n}$ with $x_i \neq x_j$ for $i \neq j$. Let $\varphi : \mathbb{R}^k \rightarrow \mathbb{R}^m$ be a feature map. If the set $\varphi(x_i)_{i=1,\ldots,n}$ is linearly independent, then the points $\varphi(x_i)_{i=1,\ldots,n}$ are linearly separable.

Lemma

If we choose $m > n$ large enough, we can always find a map φ.

- Think about our gender problem again

Non-linear SVM: Kernel tricks

- The Kernel tricks
  - Consider a transformation:

$$\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^3,\ \phi(x_1, x_2) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)'$$

  - Linear functions in feature space R^3 are quadratic functions in input space R^2:

$$g(\mathbf{x}) = w_{11}x_1^2 + w_{22}x_2^2 + w_{12}\sqrt{2}x_1x_2$$

  - Inner products in feature space R^3 can be expressed as **functions of inner products** in input space R^2

$$\begin{aligned}\langle\phi(\mathbf{x}), \phi(\mathbf{w})\rangle &= \langle(x_1^2, x_2^2, \sqrt{2}x_1x_2), (w_1^2, w_2^2, \sqrt{2}w_1w_2)\rangle \\ &= x_1^2w_1^2 + x_2^2w_2^2 + 2x_1w_1x_2w_2 \\ &= (x_1w_1 + x_2w_2)^2 \\ &= \langle\mathbf{x}, \mathbf{w}\rangle^2\end{aligned}$$

  - These functions are called **kernels**.

Non-linear SVM: Kernel tricks

- Discussion
  - (Generalized) linear classification with SVMs
    - Conceptually simple, but powerful by using kernels
  - Kernels are at the same time
    - similarity measures between arbitrary objects
    - inner products in a (hidden) feature space
  - Kerneization is implicit application of a feature map
    - The method can become non-linear in the original data
    - The method is still linear in some feature space
  - We can build new kernels from
    - Explicit inner products
    - Distances
    - Existing kernels

Non-linear SVM: Kernel tricks

- Finally, we get standard Support Vector Machine

$$
\begin{aligned}
\min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l} \xi_i \\
\text{subject to} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\
& \xi_i \geq 0, \ i = 1, \ldots, l.
\end{aligned}
$$

Outline

- Ice breaker
- Geometric viewpoint of (linear) SVM
- Statistical learning viewpoint of (linear) SVM
- Non-linear SVM: Kernel tricks
- Engineering viewpoint of SVM: practical guide
- Summary

Engineering viewpoint of SVM: practical guide

- Where to start?
  - Demo: http://www.csie.ntu.edu.tw/~cjlin/libsvm/
- When to use SVM?
  - Binary classification
- How to use SVM in practice?
  - Data preprocessing (e.g., data scaling)
  - Choose linear/nonlinear SVM
  - Choose kernel
  - Choose C (cross validation)

Outline

- Ice breaker
- Geometric viewpoint of (linear) SVM
- Statistical learning viewpoint of (linear) SVM
- Non-linear SVM: Kernel tricks
- Engineering viewpoint of SVM: practical guide
- Summary

Summary

- Machine learning = Statistics (modeling) + Optimization (solver)
- Support Vector Machine
  - Geometric viewpoint: The soft maximum margin solution for a linear classifier
  - Statistical viewpoint: Hinge loss + L2 norm regularization
  - The “Kernel trick”: a method of expanding pup from a linear classifier to a non-linear one in an efficient manner

One more thing

- Topic we do not cover today
  - SVM solvers
  - Multiclass SVM (1 vs 1, 1 vs rest, DAG)
  - Oneclass SVM (for ranking)
  - Support Vector Regression
  - Probability output for SVM
  - structured SVM
  - Multiple Kernel Learning
  - Feature mapping, i.e., approximate linear SVM to non-linear SVM
  - … …

Learning is fun; fun to learn

Queen's University Belfast

Thank you

Introduction to Support Vector Machine

Yang Hua

Mar. 2017